INTRODUCTION TO DATA ANALYTICS AND STATISTICAL ANALYSIS PROJECT

JANE MBURU

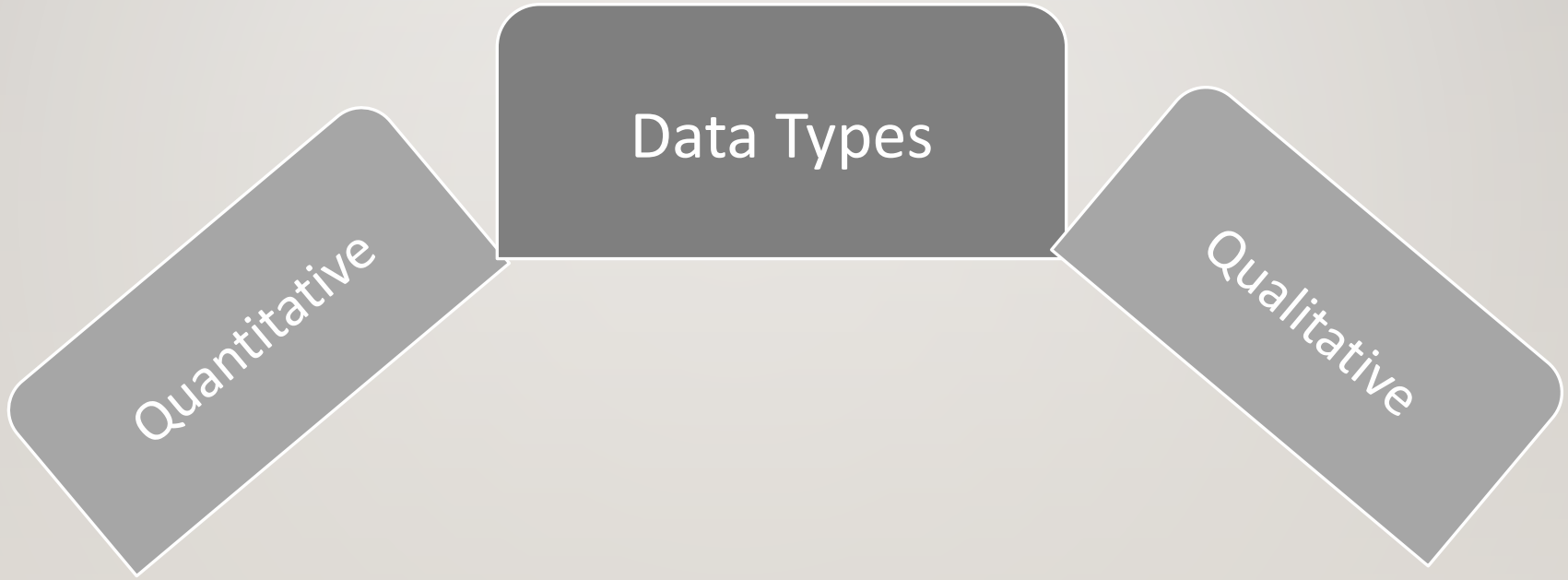
AGENDA

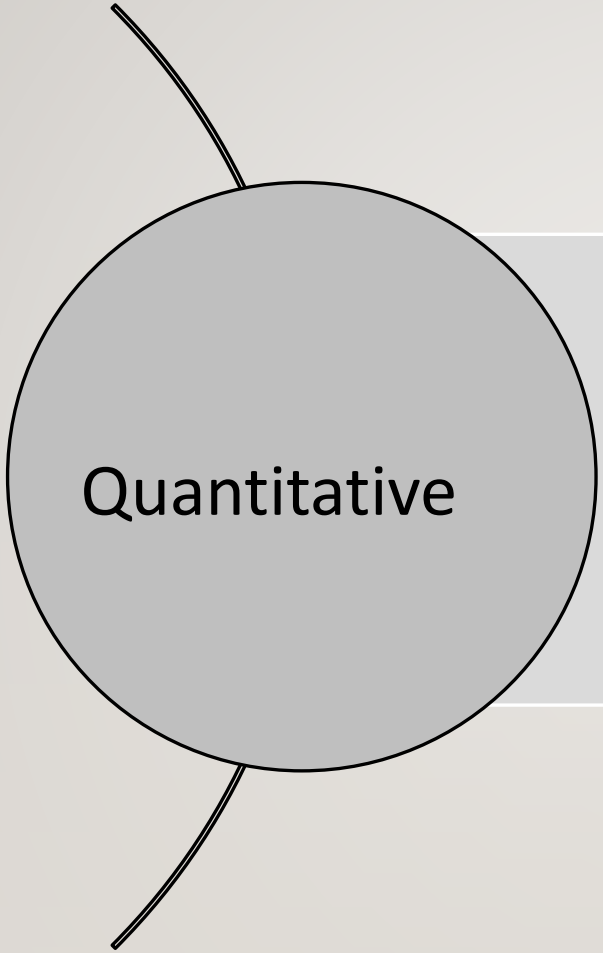
- First impressions on data
- Data types
- Data categories
- Data cleaning
- Measures of central tendency
- Hypothesis testing

FIRST IMPRESSIONS ON DATA

- Data dictionary - organized.
- Tables have unique identifiers (primary and foreign keys)
- Pizza types data table - descriptive.

DATA TYPES





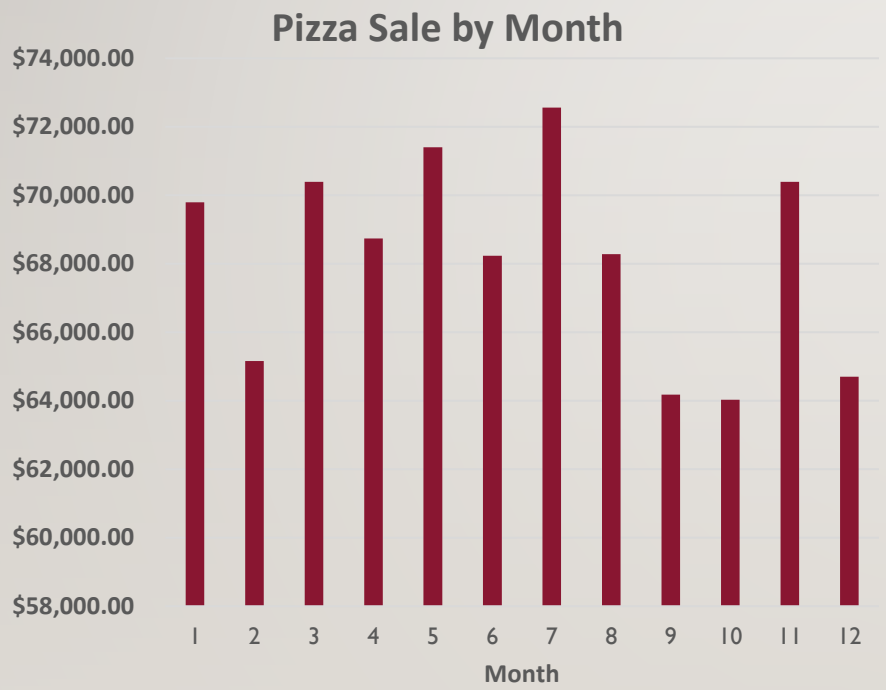
- Order_id: **Integer**
- Order_details_id: **Integer**
- Quantity: **Integer**
- Price: **Numeric (continuous)**

- Pizza_type_id: **Categorical**
- Size: **ordinal**
- Pizza_id: **Categorical**
- Name: **nominal**
- Category: **nominal**
- Ingredients: **Categorical**
- Date: **Date data type**
- Time: **Time data type**

DATA CLEANING

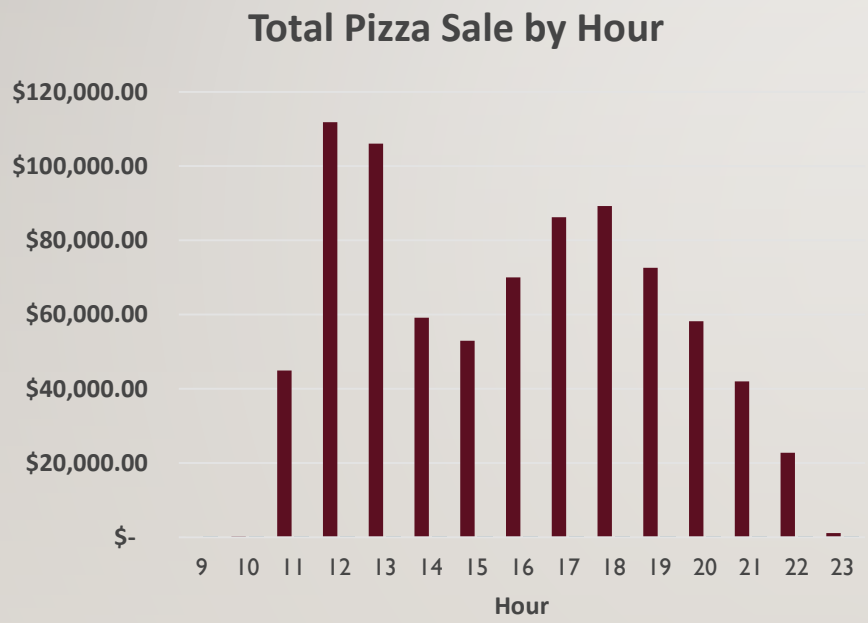
- No missing values.
- No duplicates.
- No incorrect values.
- Consistency.
- No outliers.

MEASURES OF CENTRAL TENDENCY



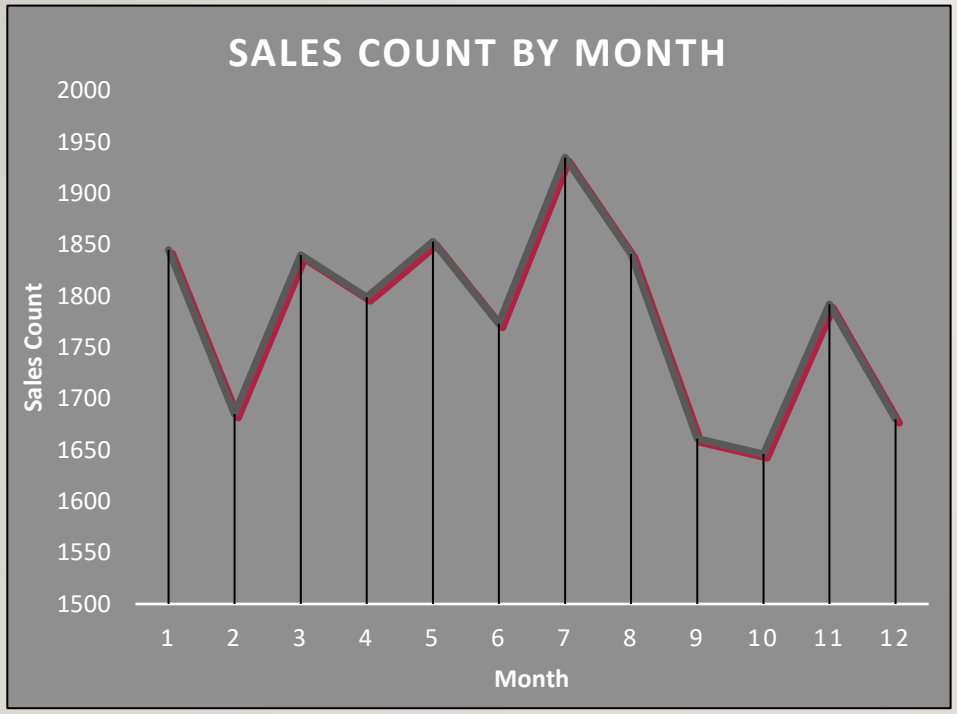
- Range = $8,530
- Std dev = $2,966
- Var = $8,796,374
- Median of $68,508 is significantly close to the mean of $68,155.
- Distribution shape is normal.
- A clustered column chart (effective visualization)

MEASURES OF CENTRAL TENDENCY



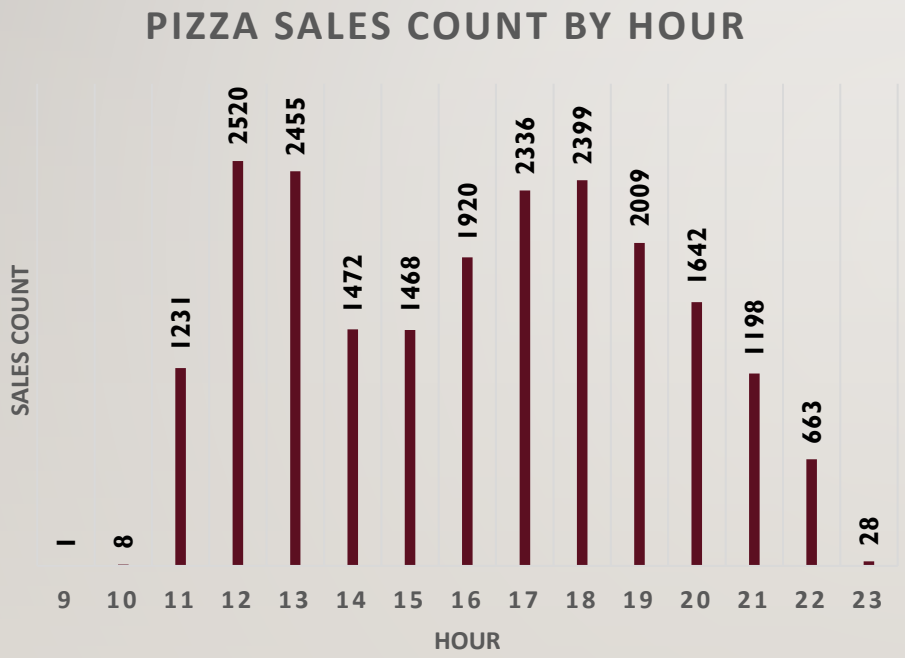
- Range = $111,795
- Std Dev = $36,622
- Var = $1,341,198,195
- Median of $58,125 is relatively close to the mean of $54,524.
- Distribution shape is bimodal.
- Visualization (clustered column chart)

MEASURES OF CENTRAL TENDENCY



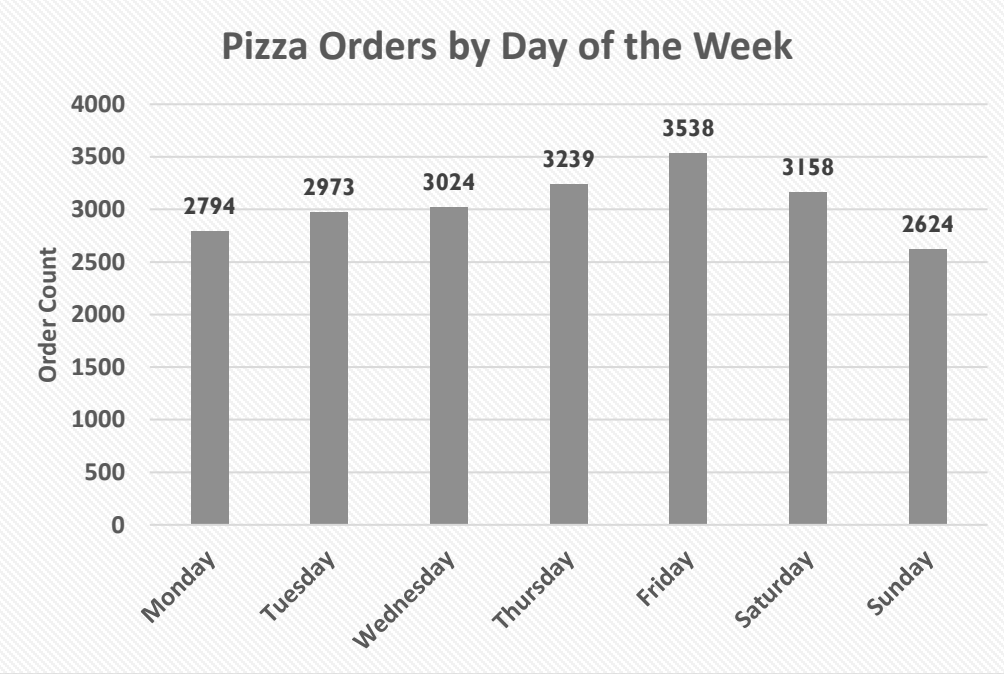
- Range = 289
- Std dev = 92
- Variance = 8,417
- The mean (1,796) and median(1,779) are relatively close.
- Distribution is almost symmetric.
- Visualization (line chart)

MEASURES OF CENTRAL TENDENCY



- Range = 2,519
- Std Dev = 900
- Variance = 809,485
- Median of 1,472 is relatively close to the mean of 1,423.
- Distribution shape is bimodal.
- Clustered column chart (visualization)

MEASURES OF CENTRAL TENDENCY



- Max= 3,538; Min= 2,624; Range=914
- Std Dev= 300; Var=89858
- Median (3,024) is significantly close to the mean of 3,050.
- The shape is approximately normal.
- Clustered column (visualization)

HYPOTHESIS TESTING…

- Null Hypothesis (**H_0)**: There is no significant difference between the average sale of pizza in the first half of the year (January to June) and the second half of the year (July to December).
- Alternate Hypothesis (**H_1)**: There is a significant difference between the average sale of pizza in the first half of the year (January to June) and the second half of the year (July to December).

SAMPLE T-TEST…

- T-test will help determine if there is a statistically significant difference between the two means:

  Mean (Jan to June) = $68,953; std dev = $2,179; sample size = 6

  Mean (July to Dec) = $67,357; std dev = $3,616; sample size = 6

- Calculations (See attached copy)

- Calculated t-value = 0.927; significance level = 0.05; degree of freedom = 10; Critical t-value from the t-distribution table = 2.228
- Since 0.927<2.228 , we fail to reject the null hypothesis. There is no significant difference between the average sale of pizza in the first half of the year and the second half of the year. The calculated t-value (0.927) is less than the critical t-value (2.228), indicating that the observed difference in means is not statistically significant at the 0.05 significance level.

KEY FINDINGS & BUSINESS IMPLICATIONS

- Sales consistency
- Peak purchase times, days, and month
- Staffing
- Inventory management
- Strategic planning

CONCLUSION

- Leveraging these data driven insights could help optimize operations, improve customer satisfaction, and increase sales.